

Review of " Exploring Hybrid Forecasting Frameworks for Subseasonal Low Flow Predictions in the European Alps" by Chang et al. The paper presents the use of a ML technique and hybrid form to improve sub-seasonal low-flow in the European alps. The results are somewhat underwhelming in the sense that the best effect of the hybrid technique requires EFAS model data, and the WR does not add information in these cases. Adding observations helped gain skill, but as the authors point out, the method then becomes a very sophisticated bias correction method. The question then arises whether similar results can be achieved by less complex methods. The study shows a major improvements in uncalibrated catchments which is useful, but why are these points not calibrated in the first place? The study is still worthwhile publishing , but I do recommend a major revision.

We thank the reviewer for the detailed review and constructive feedback. As the reviewer correctly notes, one important takeaway is that the framework outlined in this study shows the largest added value for uncalibrated catchments. These catchments are not calibrated due to systematic calibration rules within EFAS, rather than methodological limitations of this study, and we believe the findings are transferable to other non-calibrated or data-scarce regions.

We acknowledge that when observations are included, the approach might be challenged as a sophisticated bias-correction method, as discussed in the manuscript. Nevertheless, the objective of this study is to assess the performance of a hybrid forecasting framework that integrates multiple data sources under different configurations, rather than to propose a single optimal post-processing method.

We appreciate that the reviewer recognises the value of publishing the study, and the suggestions from Reviewer #1 have helped improve the manuscript. Below we provide detailed responses to the individual comments.

Major comments

1. The choice of mean flow is a very weak benchmark as it is not considering even the seasonal patterns of the streamflow, therefore making it very easy to beat and not very useful as diagnostic measure of your model performance. I strongly recommend testing the method against a benchmark of using LISFLOOD modelled with observational data, selected randomly omitting the actual year (ESP) as in Arnal et al, 2028 and Wetterhall and Di Giuseppe 2018.

We realise there is some confusion regarding the benchmark used. For the main forecast-based evaluation, we use a time-varying probabilistic climatology benchmark that does vary by season across both forecast start time and lead time. This is a very commonly accepted benchmark for forecast systems with lead times beyond a few days as outlined in Pappenberger et al. (2015), such as the case here with a forecast horizon up to 32 days.

However, the confusion relates to our description of the application of the Kling Gupta efficiency skill score (KGESS'). The KGE is arguably the most widely used metric of overall deterministic hydrological performance. Here, we apply it as an initial

screening of the 11 trained models against river discharge in reanalysis mode. Knoben et al. (2019) argued that traditionally, the Nash-Sutcliffe Efficiency (NSE) was applied widely and an NSE = 0 corresponds to using the mean flow as a benchmark predictor, but this is not the case for the KGE. A KGE = 0 does not have the same meaning as NSE = 0, despite many studies interpreting the KGE in this way. They therefore argue that papers should state more explicitly what the benchmark used should be. Therefore, we follow Knoben et al. (2019) and Harrigan et al. (2020) to rescale KGE and express it as a skill score with mean flow as a benchmark. This section will be updated in the revised manuscript with more details to more clearly explain this distinction.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979–present, *Earth System Science Data*, 12, 2043–2060, <https://doi.org/10.5194/essd-12-2043-2020>, 2020.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, *Hydrology and Earth System Sciences*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, *Journal of Hydrology*, 522, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>, 2015.

2. The selection of measures to measure skill is also not carefully considered. The authors want to provide an assessment of low flows, but have not chosen metrics that can reflect that, or modified the metrics to show their skill, for example by using the log values instead of streamflow, or selecting a sub-set of the hydrograph to focus on the low-flows. I am therefore puzzled why the study in the title says it focuses on low-flow.

This comment helped us realize that the definition of low flow events and low flow periods was not properly described in the main text of the original manuscript, which might have made it unclear to our readers that the model evaluations in the main text were based on the low flow periods only. The low flow period definition was described in Table A1 in the Appendix but we will revise the verification section 4.4 to explicitly define how low flow events are identified and how forecasts are selected for the low flow evaluation. This revision will clarify that the analyses presented in the main text focus on a subset of model outputs corresponding to low-flow periods. This conditional evaluation is consistent with the title and the primary objective of the study.

For clarification here, we did not adopt log-transformed streamflow metrics such as log-KGE, because previous studies, such as Santos et al., (2018), have highlighted important limitations of log-transformed flows. This pitfall was discussed as a

limitation in our previous study by Chang et al. (2024). To avoid this drawback in this study, we therefore focus on evaluation strategies that explicitly condition on low-flow periods, which we consider an appropriate approach for assessing low-flow prediction skill.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, *Hydrol. Earth Syst. Sci.*, 22, 4583–4591, <https://doi.org/10.5194/hess-22-4583-2018>, 2018.

Chang, A. Y.-Y., Ramos, M.-H., Harrigan, S., Prudhomme, C., Tilmant, F., Domeisen, D. I., and Zappa, M.: Exploring hydrological system performance for alpine low flows in local and continental prediction systems, *Journal of Hydrology: Regional Studies*, 56, 102 056, <https://doi.org/https://doi.org/10.1016/j.ejrh.2024.102056>, 2024.

3. The language is in general very good, but the figures are generally very difficult to interpret and need to be improved substantially. I also think that the authors sometimes show too much information rather than focusing on the important results.

We will revise the figures to improve readability and accessibility, including adjustments to colour schemes to ensure colour-blind friendliness where possible. In addition, we have revised the text to be more concise and focused on the main results.

Minor comments

1. Line 78 and onwards. The word "alpine" should not be capitalized since it is not a proper noun. The Alps are, but not alpine. Also, I find the "alpine space" a bit ambiguous, I suggest using "alpine region".

The term "Alpine Space" was originally used following its usage in several EU policy initiatives and programmes, e.g. the Interreg Alpine Space Programme. However, we recognise that this terminology may be ambiguous for a broader hydrological audience. We will therefore revise the manuscript to use the term "alpine region" throughout.

2. When discussing the study region, data and methods, it is better form to use past tense throughout, currently the paper is using both past and present tense.

We recognise that tense usage conventions differ across disciplines and journals. This study sits at the interface of hydrology, atmospheric science, and machine learning. As the present tense is commonly used when describing modelling frameworks and methods, we therefore chose to use the present tense consistently for the study region, data, and methods sections to ensure overall coherence, and will revise the manuscript to improve consistency accordingly.

3. Line 88. You use the values of streamflow expressed as mm/year. Usually, streamflow is depicted as a volume per time unit. I assume you mean specific runoff here, expressed as volume per area?

The referee is correct. Streamflow values here are expressed as area-normalized specific runoff (mm/year), not volumetric discharge. We use specific discharge to allow meaningful comparison across catchments with very different upstream areas. We will clarify the terminology and unit in the revised manuscript.

4. Figure 3 is very useful to show examples of a specific weather regime, but I think it can be improved graphically, generally by making it larger. Firstly, the colour schemes are ok, but not good for colour blind or black/white prints. Secondly, the order of the rivers in panel c are not corresponding to the anomalies, Rhine and Rhone are in the wrong order, either change the legend or the anomalies. Finally, it is very difficult to see the information in panel c. The colors are difficult to distinguish from each other, I suggest making this clearer.

Thank you for your positive and constructive feedback on Figure 3. In response, we will correct the legend in panel (c) so that the order of the rivers now corresponds correctly to the displayed anomalies (Rhine and Rhone). In addition, we will increase the overall size of the figure to improve the overall readability.

Regarding the colour scheme, Figure 3 uses a diverging colour scale to represent positive and negative anomalies, which is essential for the intended interpretation of the weather-regime signal. We acknowledge that this limits the degree to which the figure can be optimised for greyscale reproduction.

We agree that panel (c) is information-dense by design. The increased figure size and corrected legend should improve its interpretability, while preserving the full information content required for this illustrative example.

5. Line 189. I would suggest putting the reference to the LISFLOOD model documentation as a reference rather than a link in the text.

Thank you for the suggestion. The reference to the LISFLOOD model documentation will be included in the reference instead of an in-text link.

6. Line 206-256. This section needs to be rephrased and restructured, no need to put so much emphasis in the Transformer method if you use the TFT in the end for example. Also, the description of the method is too long and convoluted, I suggest moving some of this to the introduction.

We will revise Section 4.1 to be more concise by reducing the description of the general Transformer architecture and clarifying the focus on the TFT formulation used in this study. Some introductory material will be rearranged accordingly.

7. I would avoid using “/” in text, especially in headers. Choose one of the terms, or both, and write “Testing and Forecasting”

We agree and will change the section headers to avoid the use of “/”, replacing it with explicit wording (e.g. “and”).

8. Figure 4. It is almost impossible to distinguish between all stations and non-calibrated stations. I can deduct which are which since the non-calibrated always perform worse, but please select some better way to make the distinction.

We agree that it is hard to visually distinguish between the distributions for all stations and non-calibrated stations in Figure 4 in the original manuscript, particularly for models with small variability. We will revise the caption to explicitly indicate that, for each model, boxplots for all stations are always shown on the left and those for non-calibrated stations on the right. In addition, we will adjust the colour scheme to improve contrast so that the boxplots are more easily distinguishable in greyscale. We will further modify the line style of the reference KGESS' for non-calibrated stations. We believe these changes will improve the interpretability of the figure.

9. Figure 5. Also here it is very difficult to distinguish between the different experiments to make this figure useful for the reader. I understand that you want to show the skill against lead time, but you will have to show each experiment individually to help me interpret the results.

We agree that the original version of Figure 5 was difficult to interpret, particularly due to the overlap of the interquartile range (IQR) of the different experiments. To improve readability, we will remove the shaded IQR area and revise the line styles by introducing distinct line types, so the different experiments are more distinguishable. As the IQR remains important for assessing model performance, we now provide the corresponding version with shaded uncertainty bands in the Supplementary Material.

10. L379-380. I do not agree with this conclusion. If you are comparing the experiments against mean flow, any system that has high variability will be penalized by default. This does NOT mean that you can draw any conclusion on the model performance with regards to variability from this measure.

The statement referred to here is based on the forecast mode evaluation, where model performance is assessed using CRPSS relative to a time-varying climatological reference derived from observed streamflow. To avoid confusion, we will revise the manuscript.